Report on the Second NLG Challenge on Generating Instructions in Virtual Environments (GIVE-2)

Report on the *Second* Second Challenge on Generating Instructions in Virtual Environments (GIVE-2.5)

Kristina Striegnitz
Union College
striegnk@union.edu

Alexandre Denis
LORIA/CNRS
denis@loria.fr

Andrew Gargett
U.A.E. University
andrew.gargett@uaeu.ac.ae

Konstantina Garoufi
University of Potsdam
garoufi@uni-potsdam.de

Alexander Koller
University of Potsdam
akoller@uni-potsdam.de

Mariët Theune
University of Twente
m.theune@utwente.nl

Abstract

GIVE-2.5 evaluates eight natural language generation (NLG) systems that guide human users through solving a task in a virtual environment. The data is collected via the Internet, and to date, 536 interactions of subjects with one of the NLG systems have been recorded. The systems are compared using both task performance measures and subjective ratings by human users.

1 Introduction

This paper reports on the methodology and results of GIVE-2.5, the second edition of the Second Challenge on Generating Instructions in Virtual Environments (GIVE-2). GIVE is a shared task for the evaluation of natural language generation (NLG) systems, aimed at the real-time generation of instructions that guide a human user in solving a treasure-hunt task in a virtual 3D world. For the evaluation, we connect these NLG systems to users over the Internet, which makes it possible to collect large amounts of evaluation data at reasonable cost and effort.

While the shared task became more complex going from GIVE-1 to GIVE-2, we decided to maintain the same task in GIVE-2.5 (hence, the *second* second challenge). This allowed the participating research teams to learn from the results of GIVE-2 and it gave some teams (especially student teams), who were not able to participate in GIVE-2 because of timing issues, the opportunity to participate.

Eight systems are participating in GIVE-2.5. The data collection is currently underway. During July and August 2011, we collected 536 valid games, which are the basis for all results presented in this paper. This number is, so far, much lower than the number of experimental subjects in GIVE-1 and GIVE-2. Recruiting subjects has proved to be more difficult than in previous years. We discuss our hypotheses why this might be the case and hope to still increase the number of subjects during the remainder of the public evaluation period. When the evaluation period is finished, the collected data will be made available through the GIVE website.[1]

As in previous editions of GIVE, we evaluate each system both on objective measures (success rate, completion time, etc.) and subjective measures which were collected by asking the users to fill in a questionnaire. In addition to absolute objective measures, for GIVE-2.5 we also look at some new, normalized measures such as instruction rate and speed of movement. Compared to GIVE-2, we cut down the number of subjective measures and instead encouraged users to give more free-form feedback.

The paper is structured as follows. In Section 2, we give some brief background information on the GIVE Challenge. In Section 3, we present the evaluation method, including the timeline, the evaluation worlds, the participating NLG systems, and our strategy for recruiting subjects. Section 4 reports on the evaluation results based on the data that have been collected so far. Finally, we conclude and discuss future work in Section 5.

[1] http://www.give-challenge.org/research/

Figure 1: What the user sees in a GIVE world.

2 The GIVE Challenge

In GIVE, users carry out a treasure hunt in a virtual 3D world. The challenge for the NLG systems is to generate, in real time, natural language instructions that guide users to successfully complete this task.

Users participating in the GIVE evaluation start the 3D game from our website at `www.give-challenge.org`. They first download the *3D client*, the program that allows them to interact with the virtual world; they then get connected to one of the NLG systems by the *matchmaker*, which runs on the GIVE server and chooses a random NLG system and virtual world for each incoming connection. The game results are stored by the matchmaker in a database. After starting the game, the users get a brief tutorial and then enter one of three evaluation worlds, displayed in a 3D window as in Figure 1. The window shows instructions and allows the user to move around in the world and manipulate objects.

The task of the users in the GIVE world is to pick up a trophy from a safe that can be opened by pushing a sequence of buttons. Some floor tiles are alarmed, and players lose the game if they step on these tiles without deactivating the alarm first. Besides the buttons that need to be pushed, there are a number of distractor buttons that make the generation of references to target buttons more challenging. Finally, the 3D worlds contain a number of objects such as lamps and plants that do not bear on the task, but are available for use as landmarks in spatial descriptions generated by the NLG systems.

The GIVE Challenge took place for the first time in 2008–09 (Koller et al., 2010a), and for the second time in 2009–10 (Koller et al., 2010b). The GIVE-1 Challenge was a success in terms of the amount of data collected. However, while it allowed us to show that the evaluation data collected over the Internet are consistent with similar data collected in a laboratory, the instruction task was relatively simple. The users could only move through the worlds in discrete steps, and could only make 90 degree turns. This made it possible for the NLG systems to achieve a good task performance with simple instructions of the form "move three steps forward". The main novelty in GIVE-2 was that users could now move and turn freely, which made expressions like "three steps" meaningless, and made it hard to predict the precise effect of instructing a user to "turn left". Presumably due to the harder task, in combination with more complex evaluation worlds, the success rate was substantially worse in GIVE-2 than in GIVE-1. GIVE-2.5 is an opportunity to learn from the GIVE-2 experiences and improve on these results.

3 Evaluation Method

See (Koller et al., 2010a) for a detailed presentation of the GIVE data collection method. This section describes the aspects specific to GIVE-2.5, such as the timeline, the evaluation worlds, the participating NLG systems, and our strategy for recruiting subjects.

3.1 Software infrastructure

GIVE-2.5 reuses the software infrastructure from GIVE-2 described in (Koller et al., 2009) and (Koller et al., 2010b). Parts of the code were rewritten to improve how the visibility of objects is computed and how messages are sent between the components of the GIVE infrastructure: matchmaker, NLG system, and 3D client. The code is freely available at `http://code.google.com/p/give2`.

3.2 Timeline

GIVE-2.5 was first announced in July 2010. Interested research teams could start development right away, since the software interface would be the same as in GIVE-2. The participating teams had to make

their systems available for an internal evaluation period by May 23, 2011. This allowed the organizing team to verify that the NLG systems satisfied at least a minimal level of quality, while the participating research teams could make sure that their server setup worked properly, accepting connections of the matchmaker and clients to their NLG system. Furthermore, the evaluation worlds were distributed to the research teams during this period so that they could test their systems with these worlds, adapt their lexicon, if necessary, and fix any bugs that coincidentally never surfaced with the development worlds. Of course, the teams were not allowed to manually tune their systems to the new evaluation worlds in ad-hoc ways. One team had built a system that learns how to give instructions from a corpus of human-human interactions. This team was given permission to use the evaluation worlds during the internal evaluation period to collect such a corpus.

The original plan was to launch the public evaluation on June 6th. Unfortunately, some problems with the newly reworked networking code delayed the start of the public evaluation period until June 21st. At the time of writing, the public evaluation is still ongoing so that all results presented below are based on a snapshot of the data collected by August 29, 2011.

3.3 Evaluation worlds

Figure 2 shows the three virtual worlds we used in the GIVE-2.5 evaluation. The worlds were designed to be similar in complexity to the GIVE-2 worlds, and as in previous rounds of GIVE, they pose different challenges to the NLG systems. World 1 has a simple layout and buttons are arranged in ways that make it easy to uniquely identify buttons. World 2 provides challenges for the systems' referring expression generation capabilities. It contains many clusters of buttons of the same color and provides the opportunity to refer to rooms using their color or furniture. World 3 focuses on navigation instructions. One part of the world features a maze-like layout, another room contains multiple alarm tiles that the player needs to navigate around, whereas a third room has several doors and many plants but only a few other objects, making it hard for the players to orient themselves.

3.4 NLG systems

Eight NLG systems were submitted (one more than in GIVE-2, three more than in GIVE-1).

A University of Aberdeen (Duncan and van Deemter, 2011)

B University of Bremen (Dethlefs, 2011)

C Universidad Nacional de Córdoba (Racca et al., 2011)

CL Universidad Nacional de Córdoba and LORIA/CNRS (Benotti and Denis, 2011)

L LORIA/CNRS (Denis, 2011)

P1 and **P2** University of Potsdam (Garoufi and Koller, 2011)

T University of Twente (Akkersdijk et al., 2011)

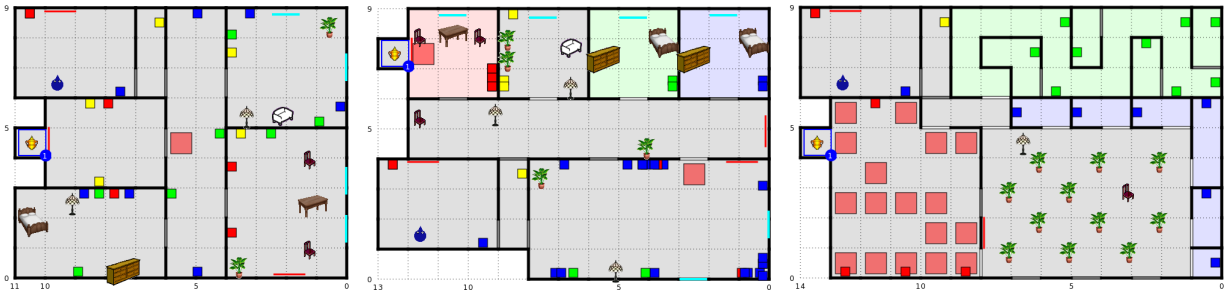
Compared to the previous GIVE editions, these systems employ more varied approaches and are better grounded in the existing CL and NLG literature. Systems A, C, L, and T are rule-based systems using hand-designed strategies. System A focuses on user engagement, T and C both focus on giving appropriate feedback to the user with C implementing the grounding model of Traum (1999), and L uses a strategy for generating referring expressions based on the Salmon-Alt and Romary (2000) approach to modeling the salience of objects.

System B uses decision trees learned from a corpus of human interactions in the GIVE domain (Gargett et al., 2010) augmented with additional annotations. System P1 uses the same corpus to learn to predict the understandability of referring expressions. The model acquired in this way is integrated into an NLG strategy based on planning. System P2 serves as a baseline for comparison against P1. Finally, system CL selects instructions from a corpus of human-human interactions in the evaluation worlds that the CL team collected during the internal evaluation phase.

See the individual system descriptions in this volume for more details about each system.

3.5 Recruiting subjects

We used a variety of avenues to recruit subjects. We posted to international and national mailing lists, gaming websites, and social networks. We had a



World 1 World 2 World 3

Figure 2: The 2011 evaluation worlds.

GIVE Facebook page and were mentioned on a relatively widely read blog. The University of Potsdam made a press release, we contributed an article to the IEEE Speech and Language Processing Technical Committee Newsletter, and submitted an entry to a list of psychological experiments online.

Unfortunately, even though we were more active in pursuing opportunities to advertise GIVE than in the last two years, we were less successful in recruiting subjects. In two months we only recorded slightly over 500 valid games, whereas in the previous years we were already well over the 1000 games mark at that point. What helped us recruit subjects in the past was that our press releases were picked up by blogs and other channels with a wide readership. Unfortunately, that did not happen this year. Maybe the summer break in the northern hemisphere, which coincided with our public evaluation phase, played a role. We are, therefore, extending the public evaluation phase into the fall, hoping to recruit enough subjects for more detailed and statistically powerful analyses than we can present in this paper.

4 Results

This section reports the results for GIVE-2.5, based on the data collected between June 21 and August 29, 2011. During this time period 536 valid games were played, that is, games in which players finished the tutorial and the game did not end prematurely due to a software or networking issue.

As in previous years, all interactions were logged. We use these logs to extract a set of objective measures. In addition, players were asked to fill in a demographic questionnaire before the game, and a questionnaire assessing their impression of the NLG system after the game. We first present some basic demographic information about our players; then we discuss the objective measures and the subjective questionnaire data. Finally, we present some further, more detailed analyses, looking at how the different evaluation worlds and demographic factors affect the results.

Again as in previous years, some of the measures are in tension with each other. For instance, a system that generates detailed and clear instructions will perhaps lead to longer games than one which tends to give instructions that are brief yet not as clear. This emphasizes that, as with previous GIVE challenges, we have aimed at a friendly challenge rather than a competition with clear winners.

4.1 Demographics

For this round of GIVE, 58% of all games were played by men and 27% by women; a further 15% did not specify their gender. While this means that we had twice as many male players as female players, we have a better gender balance than in the previous two editions of GIVE, where only about 10% of the players were female. Of all players whose IP address was geographically identifiable, about 32% were connected from Germany, 13% from the US, and 12% from the Netherlands. Argentina and France accounted for about 8% of the connections each, while 5% of them were from Sweden. The rest of the players came from 28 further countries. About half the participants (54%) were in the age range 20–29, 27% were aged 30–39, 4% were below 20, while the remaining 14% were between 40 and 69.

About 19% of the participants who answered the

task success: Did the player get the trophy?
duration: Time in seconds from the end of the tutorial until retrieval of the trophy.
distance: Distance traveled (measured in distance units of the virtual environment).
actions: Number of object manipulation actions.
instructions: Number of instructions produced by the NLG system.
words: Number of words used by the NLG system.

Figure 3: Summary of *raw objective* measures.

error rate: Number of incorrect button presses, over the total actions performed in a single game.
speed: Total distance over total time.
instruction speed: Total number of instructions over total time taken.
words per instruction: Length of instructions in number of words used.
word rate: Total number of words over total time taken.

Figure 4: Summary of *normalized objective* measures.

question were native English speakers, and an additional 73% of them self-rated their English language proficiency as at least good. The vast majority (84%) rated themselves as more experienced with computers than most people, while 47% self-rated their familiarity with 3D computer or video games as higher than that of most people. Finally, 16% indicated that they had played a GIVE game before in 2011.

4.2 Objective measures

Descriptions of the *raw* objective measures and of the *normalized* objective measures are given in Figures 3 and 4, respectively. Duration, distance travelled, and total number of actions, instructions, and words can only be compared meaningfully between games that were successful. The normalized measures, on the other hand, are independent of the result of the game. So, when comparing systems with the normalized objective measures, we have used all games in which the player managed to press at least the first button in the safe sequence.

Figures 5 and 6 show the results of raw and normalized objective measures, respectively. Task success is reported as the percentage of successfully completed games. For the other measures we give the mean value of that measure per game for each system. The figures also form groups of systems

	A	B	C	CL	L	P1	P2	T
task success	42%	32%	70%	58%	68%	66%	65%	58%
			A	A	A	A	A	A
	B			B		B	B	B
	C	C		C			C	C
duration	687	701	538	539	341	407	415	480
					A	A	A	
						B	B	B
			C	C				C
	D	D						
distance	180	204	132	153	117	128	116	166
			A		A	A	A	
			B	B		B		
	C			C				C
	D	D						D
actions	17	35	14	15	14	14	16	16
	A		A	A	A	A	A	A
		B						
instructions	165	281	254	183	211	241	235	160
	A			A				A
	B			B	B			
					C	C	C	
		D	D			D	D	
words	1894	2693	1328	1269	962	1122	1139	1024
					A	A	A	A
				B		B	B	B
			C	C		C	C	
	D							
		E						

Figure 5: Results for the *raw objective* measures.

for each evaluation measure, as indicated by the letters. If two systems do not share the same letter, the difference between these two systems is significant with $p<0.05$. Significance was tested using χ^2 for task success, and ANOVA for the other objective measures, with all systems compared pairwise using post-hoc tests (pairwise χ^2 and Tukey).

4.3 Subjective measures

Subjective measures were collected using a post-task questionnaire, which asked users to rate the instructions delivered by the NLG systems with a series of ten questions. Figure 8 shows the questions that were asked, and the average responses received. The results are based on all games, independent of success. Ratings ranged from -100 to 100, non-responses were filtered out, and, following standard practice, negative items (e.g. Q2 on confusion caused by instructions) had their scores

	A	B	C	CL	L	P1	P2	T
error rate	21%	49%	10%	11%	12%	9%	15%	19%
			A	A	A	A	A	A
	B		B	B	B		B	B
		C						
distance per sec	0.22	0.24	0.26	0.28	0.36	0.29	0.27	0.35
					A			A
				B		B		B
		C	C	C		C	C	
	D	D	D	D			D	
instructions per sec	0.21	0.36	0.48	0.32	0.62	0.56	0.54	0.33
	A							
		B		B				B
			C				C	
						D	D	
					E			
words per instruction	11.9	9.6	5.2	7.1	4.6	4.7	4.8	6.5
					A	A	A	
			B				B	
								C
				D				
		E						
	F							
words per sec	2.4	3.4	2.5	2.3	2.9	2.6	2.6	2.1
	A		A	A				A
	B		B	B		B	B	
			C		C	C	C	
		D						

Figure 6: Results for the *normalized objective* measures.

A	B	C	CL	L	P1	P2	T
Q1: Overall, the system gave me good instructions.							
-18	-31	54	24	47	31	10	-3
		A	A	A	A		
			B		B	B	
			C			C	C
D						D	D
E	E						E
Q2–10: Remaining subjective measures (summed)							
98	47	414	245	347	323	231	146
		A		A	A		
			B	B	B	B	
			C			C	C
D	D						D

Figure 7: Results for the *subjective* measures.

reversed. Once again, systems were grouped by letters where there was no significant difference between them (significance level: $p<0.05$). We used ANOVAs and post-hoc Tukey tests to test for significance.

Figure 7 furthermore shows side by side the results for the first question, which asked users for their overall impression of the system, and the results for an aggregated score obtained by summing over the rest of the questions that tried to asses specific aspects of the system.

4.4 Effects of the evaluation world and demographic factors

Which NLG system subjects interacted with is not the only factor that affects their success rate. The evaluation worlds as well as some demographic factors also had statistically significant effects.

Not surprisingly, the evaluation world affects task success ($p<0.001$), with performance in worlds 1 and 2 around 67%, but much lower in world 3 (41%). Many systems reflect the same overall pattern in their task success rates, but individual systems behave very differently as shown in Figure 9. For example, systems A and P2 do much better in world 2 than world 1, while system B does much worse in world 2 than world 1. And while all other systems have their lowest success rate in world 3, system A is doing much better in worlds 2 and 3 than in world 1.

Male players have a somewhat higher task success rate than female players (65% vs. 54%). This difference is not statistically significant, but it is close ($p=0.052$). Unfortunately, we don't have enough data, yet, to do a by system analysis of the effects that demographic properties have on task success.

The results also indicate that proficiency in English affects task success ($p=0.047$). This overall significance is due to the task success rate of subjects who rate themselves as *near native* being, with 74%, much higher than the task success rate of subjects who think of themselves as merely *good* (58%), or *very good* (57%). *Native* English speakers have a task success rate of 65%, which in pairwise comparisons is not significantly different from any of the other groups. Subjects rated their English proficiency on a 5-point scale. However, we had to drop the lowest category (*basic*) due to data scarcity.

Finally, there were effects for both familiarity with video games ($p<0.005$), and computer expertise ($p<0.05$). The questionnaire asked sub-

A	B	C	CL	L	P1	P2	T
Q1: Overall, the system gave me good instructions.							
-18	-31	54	24	47	31	10	-3
		A	A	A	A		
			B		B	B	
			C			C	C
D						D	D
E	E						E
Q2: I was confused about which direction to go in.							
-22	-16	52	27	31	26	16	-17
		A	A	A	A		
			B	B	B	B	
C	C						C
Q3: I could easily identify the buttons the system described to me.							
37	3	60	46	42	39	16	23
A		A	A	A	A		
B			B	B	B	B	B
C	C					C	C
Q4: I had to re-read instructions to understand what I needed to do.							
14	-4	50	19	53	19	1	2
		A		A			
		B	B		B		
C	C		C		C	C	C
Q5: The system's instructions were visible long enough for me to read them.							
-10	-12	42	13	51	37	38	24
		A		A	A	A	A
		B	B		B	B	B
C	C		C				C

A	B	C	CL	L	P1	P2	T
Q6: The system's instructions came too late or too early.							
-6	-10	36	-3	34	24	19	2
		A		A	A	A	
B			B		B	B	B
C	C		C			C	C
Q7: The system immediately offered help when I was in trouble.							
-13	1	52	17	38	48	35	1
		A		A	A	A	
			B	B		B	
	C		C				C
D	D						D
Q8: The system gave me useful feedback about my progress.							
-4	-16	62	37	23	57	33	27
		A	A		A	A	
			B		B	B	B
			C	C		C	C
D	D						
Q9: The system was very friendly.							
25	31	54	46	49	54	42	35
	A	A	A	A	A	A	A
B	B	B	B	B		B	B
Q10: I felt I could trust the system's instructions.							
0	-25	69	38	52	44	30	12
		A	A	A	A		
			B	B	B	B	
			C			C	C
D						D	D
E	E						

Figure 8: Results for individual questionnaire items.

jects to rate themselves as being much less familiar with video games/experienced with computers than most people, less familiar/experienced than most people, equally familiar/experienced, more familiar/experienced, or much more familiar/experienced. Again, due to data scarcity, we had to collapse the lowest two and highest two categories for familiarity with video games and the lowest three categories for computer expertise. On closer inspection, these overall significant effects are accounted for by a significant difference in task success ($p<0.001$) between players who rated themselves as *less familiar* with video games than most people (51% task success rate) and players who rated themselves as *more familiar* (69%). Similarly, the subjects who think of themselves as *much more* experienced with computers than most people (66%) are significantly more successful than subjects who think they are *less or equally* experienced than most people (49%).

4.5 Discussion

The objective and subjective measures largely agree in ranking systems C, CL, L, P1, P2, T before systems A and B. The first six systems do not differ significantly from each other in terms of task success or error rate. However, there are some significant differences between them when looking at the other objective measures. For example, games with

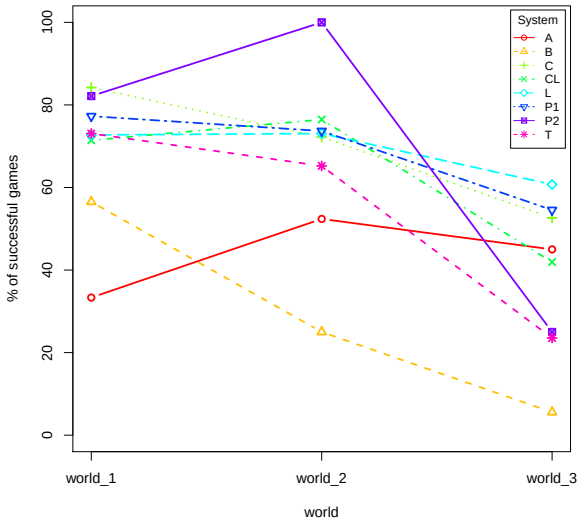


Figure 9: Effect of the different evaluation worlds on the task success rate of the NLG systems.

systems L, P1, and P2 are shorter than than those with systems C and CL, while system T is sitting in between the two groups.

Interestingly, shorter durations do not necessarily coincide with the players moving faster. For instance, players interacting with systems P1 and P2 move significantly slower than players who interact with system L. System L also delivers its instructions at a very fast pace, followed by systems P1 and P2. Those are the same systems that achieve the shortest game durations, and they also make the group of systems which produces the most concise instructions. However, it is not necessary for an NLG system to be as fast paced as the L and P systems to be successful. If we compare the two systems with the highest task success rates, systems C (70%) and L (68%), we see that L has very short games, fast moving players, and delivers its concise instructions at an extremely high rate. C, on the other hand, yields significantly longer games, has players that move at a significantly slower speed, and produces significantly longer instructions (though still concise compared to some other systems) at a much lower rate.

There is also some indication, though, that being too slow and wordy might be detrimental. Systems A and B, the least effective in terms of task success and error rate, have extremely long games, slow players, and long instructions that get sent at a slow pace.

As mentioned above, the subjective measures largely agree with the ranking suggested by the objective measures: systems C, CL, L, P1, P2, T are ranked before systems A and B. However, the top group is a little more split up. Systems C, L, and P1 are ranked highest both by Q1, the questionnaire item asking for an overall assessment, and by the summed scores for the remaining questionnaire items. Systems CL and P2, on the other hand, come in the next tier according to these subjective measures, while system T follows.

System C is doing well on questionnaire items that have to do with timing (such as Q6 and Q7), suggesting that even though it is slower than some of the other most successful systems, its instructions are well-timed. One interesting point to notice is that system A, which overall is not so successful, is doing relatively well on item Q3. In fact, referring expression generation is one of the aspects system A's team focused on.

Comparing this year's results to those of GIVE-2, we can report that task success has increased somewhat. The task success rate of systems in GIVE-2 ranged from 3% to 47% with a mean success rate of 29%. For GIVE-2.5, task success rates range from 32% to 70% with a mean of 57%. Though these results are measured in different worlds and are thus not directly comparable, they do provide some evidence of the overall increasing quality of systems entered in this round of GIVE.

Interestingly, the overall quality ratings (Q1) did not go up across the board in a similar way, although the systems that did best on this measure in GIVE-2.5 had somewhat higher scores than the best systems in the previous installment of GIVE. In GIVE-2, the systems had a mean score for that question that ranged from -33 to 36. In GIVE-2.5, the mean scores ranged from -31 to 54. Some of the other subjective measures improved more dramatically, though. For example, the systems' mean ratings for Q2 (*I was confused which direction to go in*) ranged from -32 to 21 in GIVE-2, but from -22 to 52 in GIVE-2.5.

Unfortunately, we don't have enough data, yet, to compare the effect that demographic factors have on

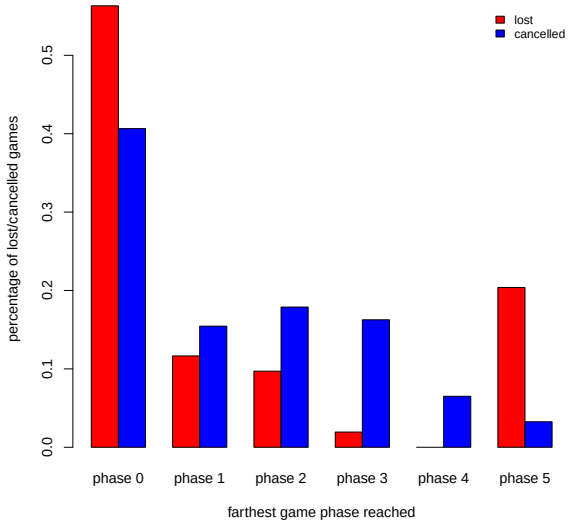


Figure 10: Player progress before they lose/cancel.

individual systems. By the end of our evaluation period, we will hopefully be able to make that analysis.

5 Conclusions and Outlook

This paper has described the methodology and results of GIVE-2.5, the second edition of the Second Challenge on Generating Instructions in Virtual Environments. In a number of ways, GIVE-2.5 expanded successfully on GIVE-2. Eight NLG systems participated in GIVE-2.5, one more than in GIVE-2. These systems represent a broader variety of approaches to NLG than seen before in a GIVE challenge, and the instructions they generate are of a higher quality.

Unexpectedly, our efforts to recruit subjects over the Internet were not as successful as in previous years. We think that this is mostly due to less luck with getting our advertising into channels that reach a broad audience, which was possibly exacerbated by the timing of the public evaluation period during the northern hemisphere summer break. It would be desirable to develop an advertising strategy for future editions of the challenge that can distribute our call to play GIVE more reliably.

One problem we already identified in GIVE-1 and GIVE-2 is that the task is not as engaging for players as modern 3D games are. As in GIVE-2, this is evidenced by the observation that many players cancel or lose the game before they ever press the first button in the safe sequence. (Figure 10 shows how close subjects got to finding the trophy before losing or canceling. Phase 0 means that not even the first button of the safe sequence was pressed successfully; phase 1 means that one button of the safe sequence was pressed successfully, etc.) The free text comments also contain complaints in that direction. We did not expect this problem to disappear, since the task is the same as in GIVE-2, but its persistence re-confirms that the next revision of GIVE needs to address this issue.

We are currently discussing the task and timeline for GIVE-3. The plan is to make a substantial change to the task. The specification of this new task and the implementation of the necessary software infrastructure needs some time, so that we will most likely not organize another edition of GIVE before 2013. However, Oliver Lemon and Srini Janarthanam will organize a challenge similar to GIVE in 2012, called *Generating Route Instructions under Uncertainty in Virtual Environments* (GRUVE). Its main features are that the game world will be an outdoor environment based on publicly available map data, and that it will be possible for NLG systems to interact with users in a more dialog-like fashion by generating questions plus a set of possible answers for the user to choose from. In addition, there will be an *uncertainty* track, where the player coordinates sent to the NLG system by the client will be artificially distorted in order to simulate a noisy GPS signal. (See Janarthanam and Lemon (2011) in this volume for more details.) We encourage everybody interested in GIVE to consider participating in GRUVE.

Acknowledgements

We would like to thank all participating research teams for their contributions, and the GIVE Steering Committee (Donna Byron, Justine Cassell, Robert Dale, Johanna Moore, Jon Oberlander) for their help and support in organizing the GIVE challenges.

References

S. Akkersdijk, M. Langenbach, F. Loch, and M. Theune. 2011. The Thumbs Up! Twente system for GIVE 2.5.

In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, Nancy, France.

L. Benotti and A. Denis. 2011. CL system: Giving instructions by corpus based selection. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, Nancy, France.

A. Denis. 2011. The Loria instruction generation system L in GIVE 2.5. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, Nancy, France.

N. Dethlefs. 2011. The Bremen system for the GIVE-2.5 Challenge. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, Nancy, France.

B. Duncan and K. van Deemter. 2011. Direction giving: an attempt to increase user engagement. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, Nancy, France.

A. Gargett, K. Garoufi, A. Koller, and K. Striegnitz. 2010. The GIVE-2 corpus of Giving Instructions in Virtual Environments. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Malta.

K. Garoufi and A. Koller. 2011. The Potsdam NLG systems at the GIVE-2.5 Challenge. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, Nancy, France.

S. Janarthanam and O. Lemon. 2011. The GRUVE Challenge: Generating Routes under Uncertainty in Virtual Environments. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, Nancy, France.

A. Koller, D. Byron, J. Cassell, R. Dale, J. Moore, J. Oberlander, and K. Striegnitz. 2009. The software architecture for the First Challenge on Generating Instructions in Virtual Environments. In *Proceedings of the EACL-09 Demo Session*.

A. Koller, K. Striegnitz, D. Byron, J. Cassell, R. Dale, J. Moore, and J. Oberlander. 2010a. The First Challenge on Generating Instructions in Virtual Environments. In E. Krahmer and M. Theune, editors, *Empirical Methods in Natural Language Generation*, volume 5790 of *LNCS*, pages 337–361. Springer.

A. Koller, K. Striegnitz, A. Gargett, D. Byron, J. Cassell, R. Dale, J. Moore, and J. Oberlander. 2010b. Report on the Second NLG Challenge on Generating Instructions in Virtual Environments (GIVE-2). In *Proceedings of the Generation Challenges Session at the 6th International Natural Language Generation Conference*, Trim, Ireland.

D.N. Racca, L. Benotti, and P. Duboue. 2011. The GIVE-2.5 C generation system. In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation*, Nancy, France.

S. Salmon-Alt and L. Romary. 2000. Generating referring expressions in multimodal contexts. In *Proceedings of the Workshop on Coherence in Generated Multimedia (Co-located with INLG)*, Mitzpe Ramon, Israel.

D.R. Traum. 1999. Computational models of grounding in collaborative systems. In *Working Notes of the AAAI Fall Symposium on Psychological Models of Communication*, North Falmouth, MA, USA.